**Title:** Temporal patterns in the evolutionary genetic distance of SARS-CoV-2 during the COVID-19 pandemic

**Running Title**: Mutation accumulation in SARS-CoV-2

**Keywords:** SARS-CoV-2; genetic distance; whole genome sequencing

**Authors**:

Jingzhi Lou[1,+], Shi Zhao[1,2,+], Lirong Cao[1,+], Zigui Chen[3], Renee WY Chan[4], Marc KC Chong[1,2], Benny CY Zee[1,2], Paul KS Chan[3], and Maggie H Wang[1,2*]

**Author affiliation:**

**1** JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China

**2** CUHK Shenzhen Research Institute, Shenzhen, China

**3** Department of Microbiology, the Chinese University of Hong Kong, Hong Kong SAR, China

**4** Department of Paediatric, the Chinese University of Hong Kong, Hong Kong SAR, China

+ Joint first authors.

***** Correspondence to: maggiew@cuhk.edu.hk (MHW)

**Email addresses of all authors**

JL: jzlou@qq.com

SZ: zhaoshi.cmsa@gmail.com

LC: caolr@link.cuhk.edu.hk

ZC: zigui.chen@cuhk.edu.hk

RWYC: reneewy@cuhk.edu.hk

MKCC: marc@cuhk.edu.hk

BCYZ: bzee@cuhk.edu.hk

PKSC: paulkschan@cuhk.edu.hk

MHW: maggiew@cuhk.edu.hk

**Abstract:**

**Background:** During the pandemic of coronavirus disease 2019 (COVID-19), the genetic mutations occurred in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cumulatively or sporadically. In this study, we employed a computational approach to identify and trace the emerging patterns of the SARS-CoV-2 mutations, and quantify accumulative genetic distance across different periods and proteins.

**Methods:** Full-length human SARS-CoV-2 strains in United Kingdom were collected. We investigated the temporal variation in the evolutionary genetic distance defined by the Hamming distance since the start of COVID-19 pandemic.

**Findings:** Our results showed that the SARS-CoV-2 was in the process of continuous evolution, mainly involved in spike protein (S protein), the RNA-dependent RNA polymerase (RdRp) region of open reading frame 1 (ORF1) and nucleocapsid protein (N protein). By contrast, mutations in other proteins were sporadic and genetic distance to the initial sequenced strain did not show an increasing trend.

## Introduction:

The pandemic of coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) poses severe threat to public health globally. The genetic mutations in SARS-CoV-2 have been detected frequently. Although most mutations occur sporadically and are purged shortly, it appears that some mutations gradually reach fixation [1]. Given increasing numbers of fixed mutations, the circulating SARS-CoV-2 strains may diverge from the original strain in terms of an increasing accumulative genetic distance. The accumulation of genetic distance reflects a steady viral evolutionary process that may affect the characteristics of SARS-CoV-2, immune recognition, and effectiveness of antivirals targeting the pathogen. In this study, we employed a computational approach to identify and trace the emerging patterns of the SARS-CoV-2 mutations, and quantify accumulative genetic distance across different periods and proteins.

## Data and methods:

The full-length human SARS-CoV-2 strains in United Kingdom were obtained from Global Initiative on Sharing all Influenza Data (GISAID) [2] on September 23, 2020. A total of 40,527 strains were collected with the collection date ranging from January 27 to September 14, 2020. We used a stratified sampling scheme to randomly selected sequences in biweekly time interval. See Supplementary Materials S1 for the details of the sampling scheme and summary. As one of the first reported sequences, ‘China/Wuhan-Hu-1/2019’, which was collected in December 31, 2019, was considered as the reference strain for sequence alignment, and as the initial strain for genetic distance calculation against other strains. The genetic distance to the initial strain was defined by the Hamming distance. We investigated the temporal variation in the accumulative genetic distance in the cross-sectional series. To better observe the dynamic trend of genetic distance, a sliding window was applied (Supplementary Materials S2). Multiple sequences alignment was performed using MEGA-X (version 10.1.8).

## Results and discussions:

We observed that the genetic distance to the initial strain steadily increased with time since February 2020 (Figure 1). The overall distance of all proteins rose from 0 to a peaking level at 9.27 codons on

August 21, 2020, which implied mutations with evolutionary advantages occurred and gradually accumulated. Moreover, deleterious mutants can only preserve for a short period in virus population and disappeared due to functional issues or less adaption to environment, which explains the fluctuations in the genetic distance curve, e.g. open reading frame 8 (ORF8) in Figure 1. By further observing the distance in each protein, we found continuous increasing trends in the spike protein (S protein), the ORF1 (especially in the RNA-dependent RNA polymerase, Supplementary Materials S3) and the nucleocapsid protein (N protein), and no obvious trends in the membrane protein (M protein), the envelope protein (E protein), the ORF3, ORF6-8 and ORF10, see Figure 1. It suggested a stronger natural selection pressure on the S protein, the ORF1 and the N protein. The S protein is responsible for receptor recognition and membrane fusion, and contains a receptor binding domain (RBD), which is considered as the target for the SARS-CoV-2 vaccines under development[3-5]. The RdRp in ORF1 plays a central role in the replication and transcription cycle of SARS-CoV-2 and thus is considered as a target for nucleotide analog antiviral inhibitors[6]. N protein is a multifunctional and highly immunogenic determinant, whose function is mediating the packaging of the viral RNA genome into the nascent virion[7]. These three proteins were critical in determining the course of transmission, infection and reproduction of the virus, and thus accumulated mutations in these proteins might be related to the viral adaptation to the environment both in vivo and in vitro. By contrast, genetic distances of the proteins with sporadic mutation were usually below 0.5 codon without demonstration of an increasing trend. To date, the mutations in these proteins have not maintained, which might be due to their less competitive strength in facilitating the viral adaptation to the environment.

Our results showed that the SARS-CoV-2 was in the process of continuous evolution, mainly involved in the S protein, the RdRp region of ORF1 and the N protein. In August 24, 2020, the first cases of COVID-19 re-infection was officially identified and reported that the viruses corresponding to the first and second infections carried 13 amino acid differences [8]. Another study showed that functional S protein and N-terminal domain of SARS-CoV-2 with mutations conferred resistance to monoclonal antibodies [9]. Continuous evolution of the virus might bring considerable challenge to the development of antiviral drugs and vaccines.

**Conclusion:**

This study presents the evolutionary process of SARS-CoV-2 virus from the aspect of genetic distance. The continuous mutation accumulation was observed in genes encoding the S protein, the RdRp region and the N protein, but not be observed in genes encoding other proteins to date. Therefore, future investigation is warranted to study the characteristics and the effects associated with the accumulative mutations in SARS-CoV-2, as well as the treatment or control strategies.

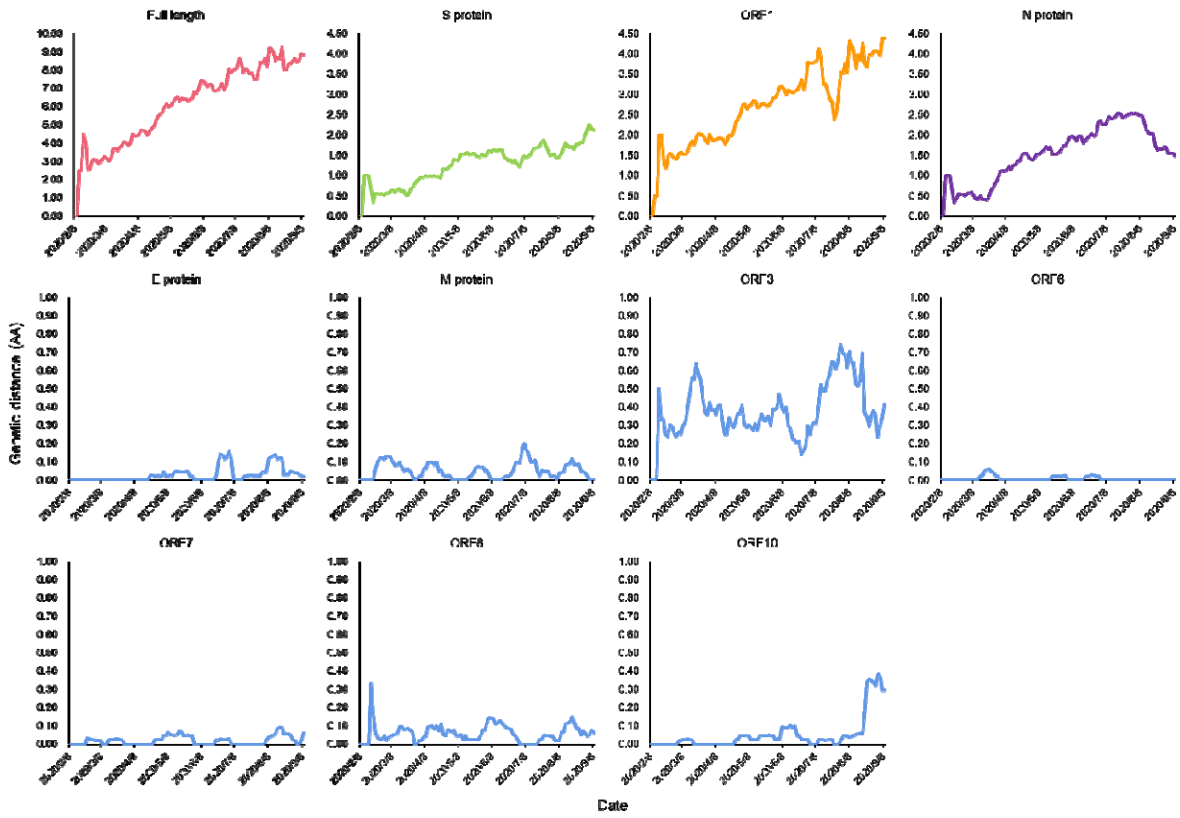


**Figure 1. The time-varying genetic distance from initial strain in different proteins.**

The green, orange and purple lines represent genetic distance in S protein, ORF1 and N protein with increasing trends, respectively. The blue lines represent genetic distance in M protein, E protein, ORF3, ORF6, ORF7, ORF8 and ORF10 without notable trends.

## Declarations

### Authors' contributions

MHW conceived the study. JL collected the data and carried out the analysis. JL, SZ, LC and MHW discussed the results. JL drafted the first manuscript. JL, SZ, LC and MHW reviewed and edited the manuscript. All authors critically read and revised the manuscript and gave final approval for publication.

### Funding

This work is supported by the Health and Medical Research Fund (HMRF) Commissioned Research on COVID-19 [COVID190103] and [INF-CUHK-1] of Hong Kong SAR, China, and partially supported by the National Natural Science Foundation of China (NSFC) [31871340] and CUHK Direct Grant [4054524].

### Disclaimer

The funding agencies had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

### Acknowledgements

The SARS-CoV-2 sequences were collected from Global Initiative on Sharing all Influenza Data (GISAID) accessible via https://www.gisaid.org/. We thank the contribution of the submitting and the originating laboratories. This study was conducted using the resources of Alibaba Cloud Intelligence High Performance Cluster computing facilities, which is made free for COVID-19 research.

### Conflict of interests

MHW is a shareholder of Beth Bioinformatics Co., Ltd. BCYZ is a shareholder of Beth Bioinformatics Co., Ltd and Health View Bioanalytics Ltd. Other authors declared no competing interests.

### Ethics approval and consent to participate

137 The human SARS-CoV-2 strains were collected via public domains, and thus neither ethical

138 approval nor individual consent was not applicable.

139 **Availability of materials**

140 All data used in this work were publicly available.

## References:

1. Benvenuto D, Demir AB, Giovanetti M, Bianchi M, Angeletti S, Pascarella S, et al. Evidence for mutations in SARS-CoV-2 Italian isolates potentially affecting virus transmission. Journal of medical virology. 2020. Epub 2020/06/04. doi: 10.1002/jmv.26104. PubMed PMID: 32492183; PubMed Central PMCID: PMCPMC7300971.

2. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2017;22(13). Epub 2017/04/07. doi: 10.2807/1560-7917.Es.2017.22.13.30494. PubMed PMID: 28382917; PubMed Central PMCID: PMCPMC5388101.

3. Jackson LA, Anderson EJ, Rouphael NG, Roberts PC, Makhene M, Coler RN, et al. An mRNA Vaccine against SARS-CoV-2 - Preliminary Report. The New England journal of medicine. 2020. Epub 2020/07/15. doi: 10.1056/NEJMoa2022483. PubMed PMID: 32663912; PubMed Central PMCID: PMCPMC7377258.

4. Mercado NB, Zahn R, Wegmann F, Loos C, Chandrashekar A, Yu J, et al. Single-shot Ad26 vaccine protects against SARS-CoV-2 in rhesus macaques. Nature. 2020. Epub 2020/07/31. doi: 10.1038/s41586-020-2607-z. PubMed PMID: 32731257.

5. Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, et al. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell. 2020;181(4):894-904.e9. Epub 2020/04/11. doi: 10.1016/j.cell.2020.03.045. PubMed PMID: 32275855; PubMed Central PMCID: PMCPMC7144619.

6. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science (New York, NY). 2020;368(6492):779-82. Epub 2020/04/12. doi: 10.1126/science.abb7498. PubMed PMID: 32277040; PubMed Central PMCID: PMCPMC7164392.

7. Zinzula L, Basquin J, Bohn S, Beck F, Klumpe S, Pfeifer G, et al. High-resolution structure and biophysical characterization of the nucleocapsid phosphoprotein dimerization domain from the Covid-19 severe acute respiratory syndrome coronavirus 2. Biochemical and Biophysical Research Communications. 2020. Epub 2020/10/12. doi: 10.1016/j.bbrc.2020.09.131. PubMed PMID: 33039147; PubMed Central PMCID: PMCPMC7532810.

8. To KK, Hung IF, Ip JD, Chu AW, Chan WM, Tam AR, et al. COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2020. Epub 2020/08/26. doi: 10.1093/cid/ciaa1275. PubMed PMID: 32840608; PubMed Central PMCID: PMCPMC7499500.

9. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JCC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. 2020:2020.07.21.214759. doi: 10.1101/2020.07.21.214759 %J bioRxiv.